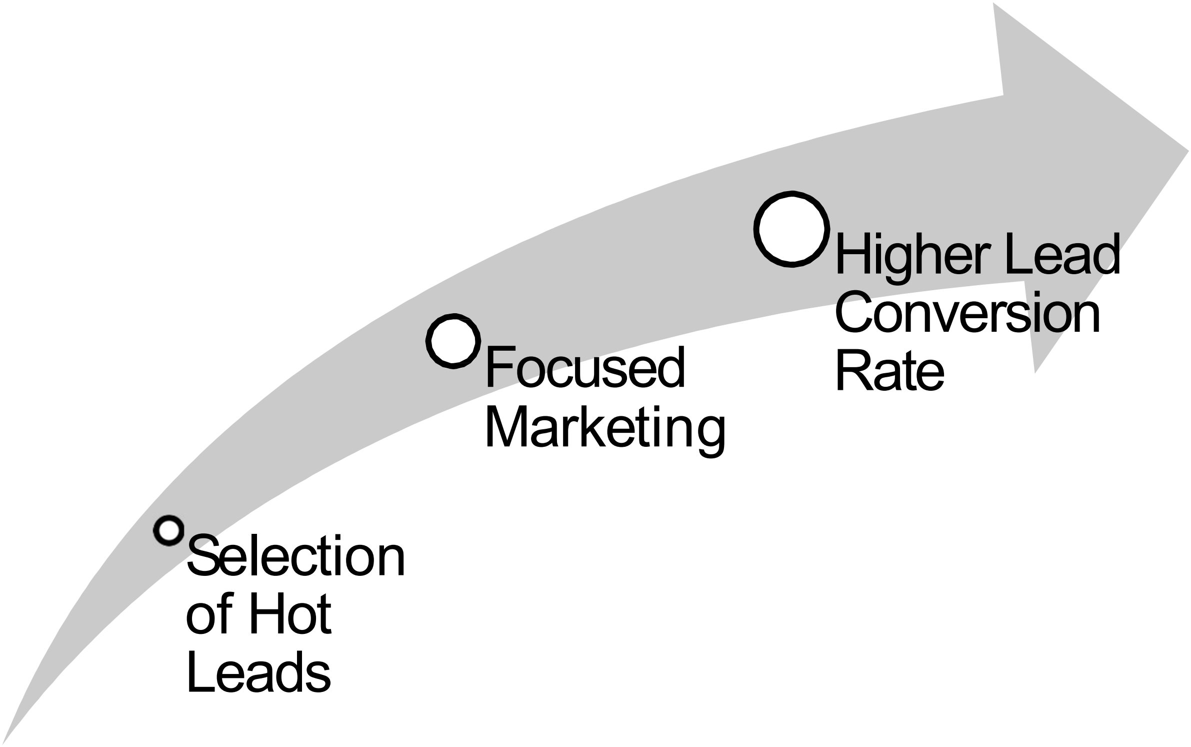
LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Business Objective

Tohelp XEducation select most promising leads ***(Hot Leads)***, i.e. the leads that are most likely to convert into paying customers.



Selection of Hot Leads

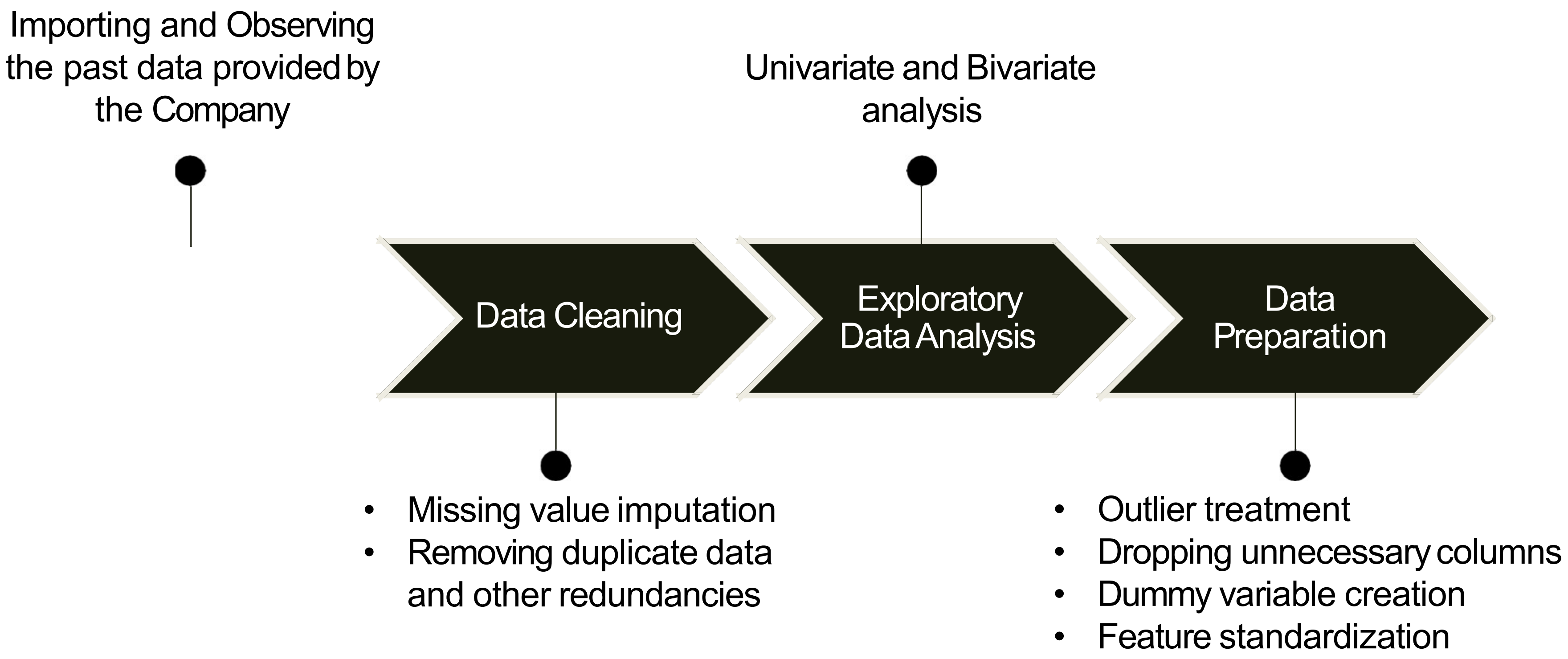
Focused Marketing

Higher Lead Conversion Rate

METHODOLOGY

Tobuild a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.

Target Lead Conversion Rate ≈ 80%



Importing and Observing the past data provided by the Company

Data Cleaning

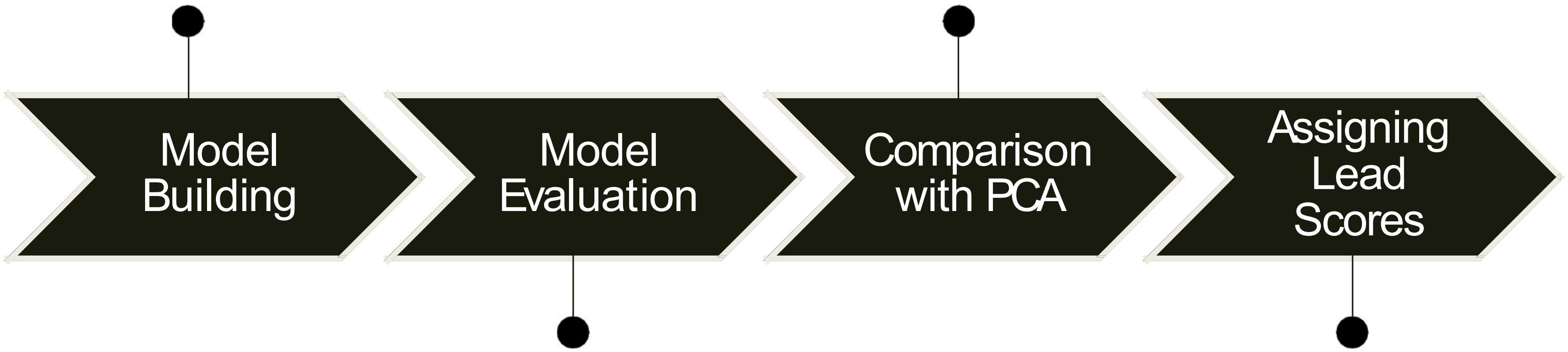
- Missing value imputation
- Removing duplicate data and other redundancies

Exploratory Data Analysis

Univariate and Bivariate analysis

Data Preparation

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization



Model Building

- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

Model Evaluation

- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

Comparison with PCA

- Building another model using PCA
- Comparing the two models

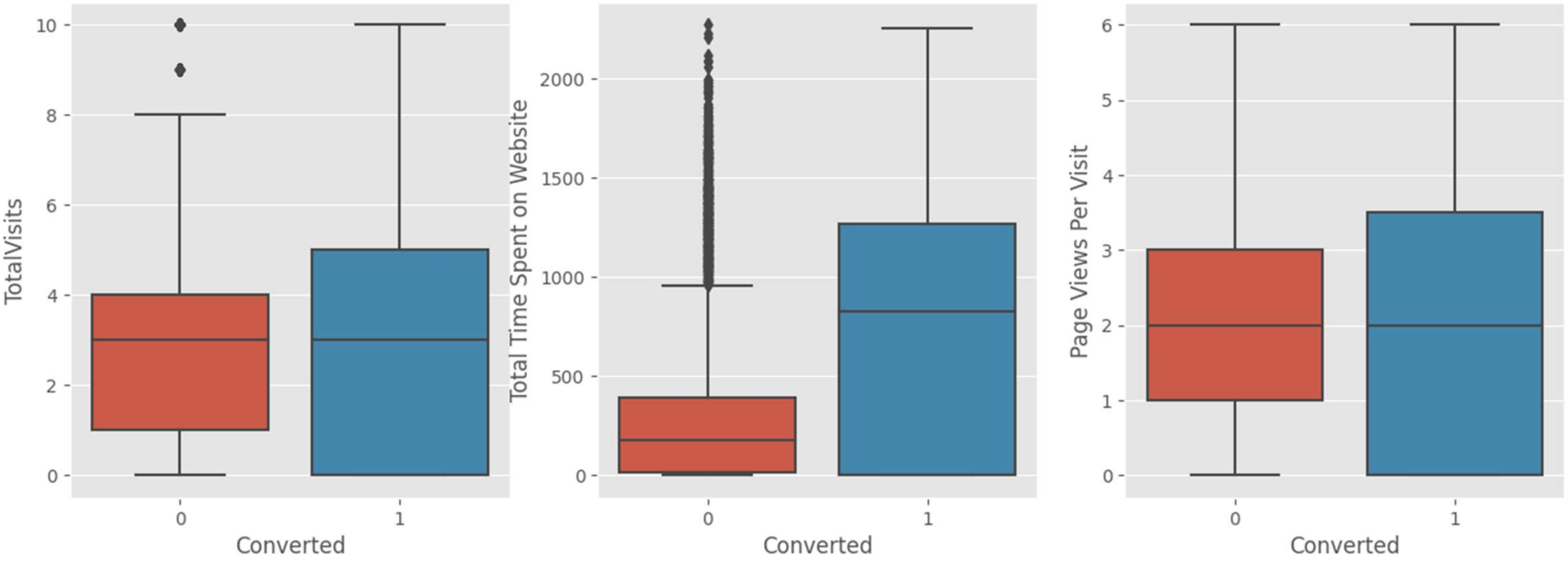
Assigning Lead Scores

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
  Lead Score = Probability * 100

DATA VISUALIZATION

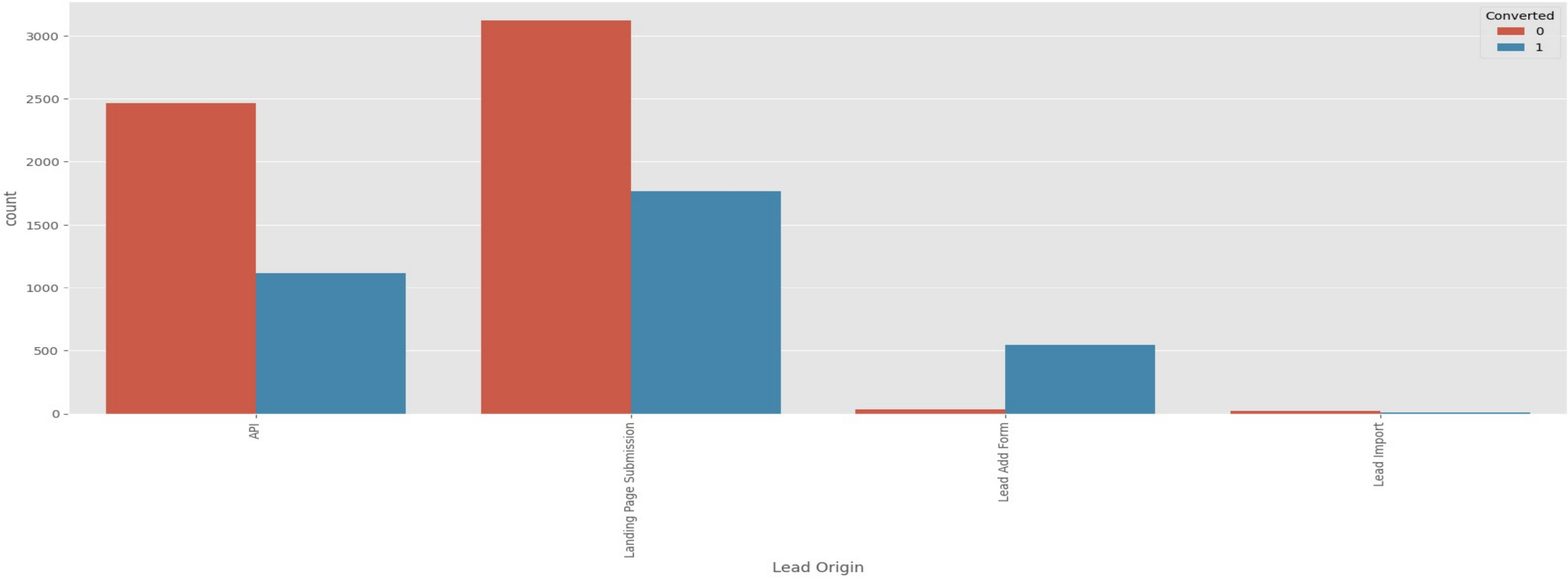
- To identify important features
- To get insights

Numerical Variables



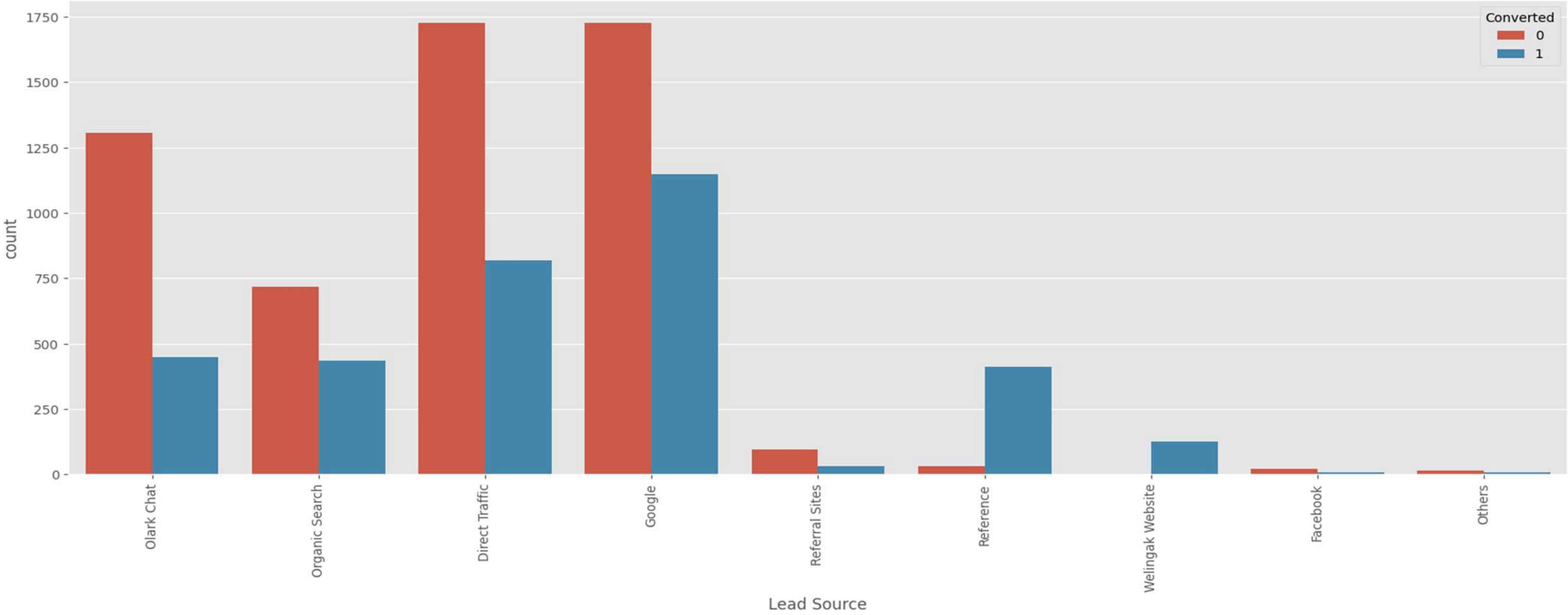
People spending more time on website are more likely to get converted.

Lead Origin



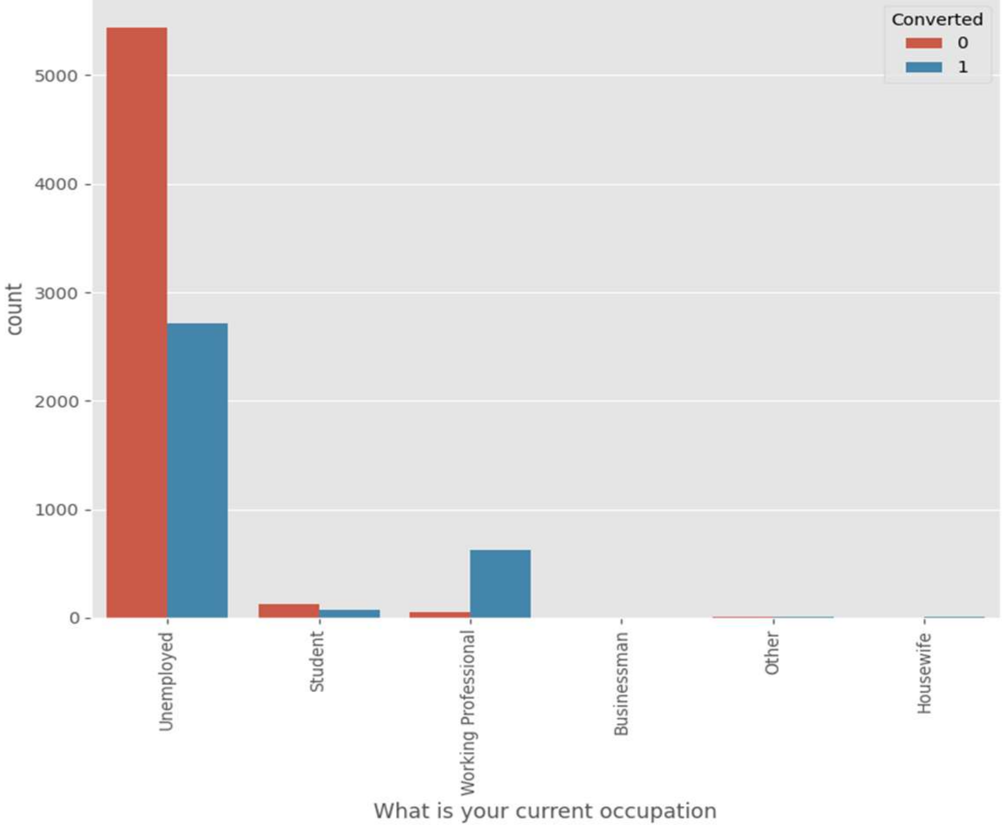
- ***'API'*** and '***Landing Page Submission***' generate the most leads but have less conversion rates, whereas '***Lead AddForm***' generates less leads but conversion rate is great.
- **Tryto increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'**.

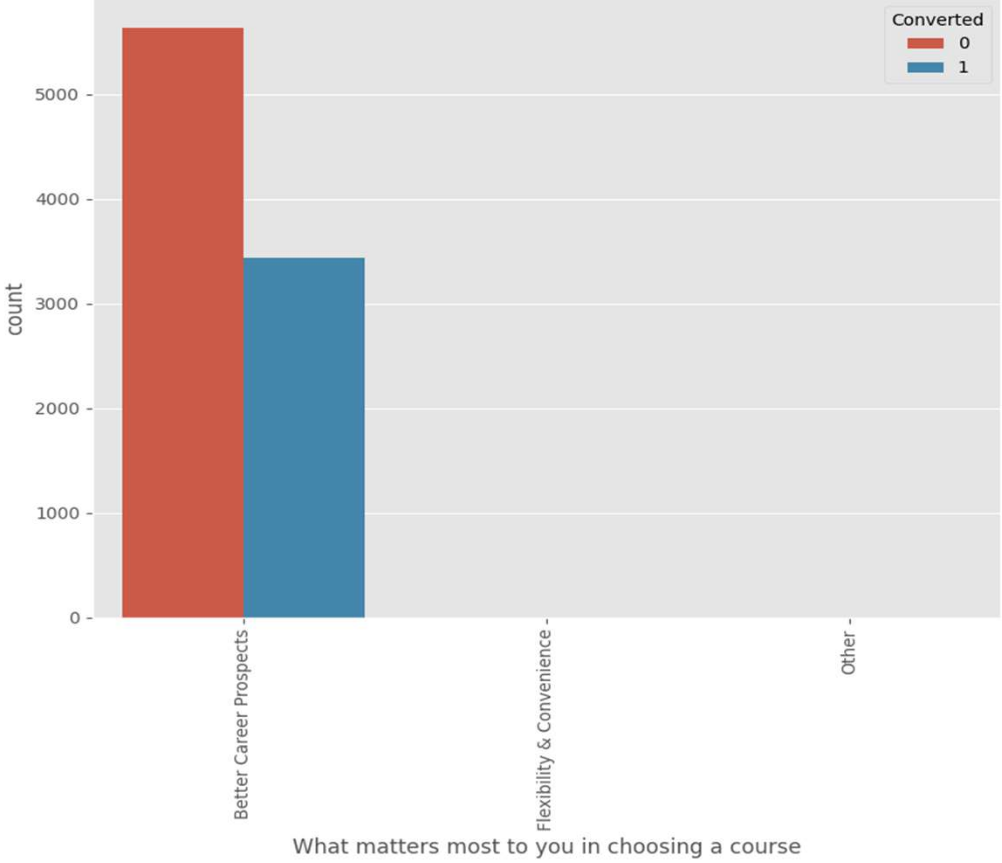
Lead Source



- Very high conversion rates for lead sources ‘**Reference**’ and ‘**Welingak Website’.**
- Most leads are generated through **‘Direct Traffic**’ and ‘**Google**’.

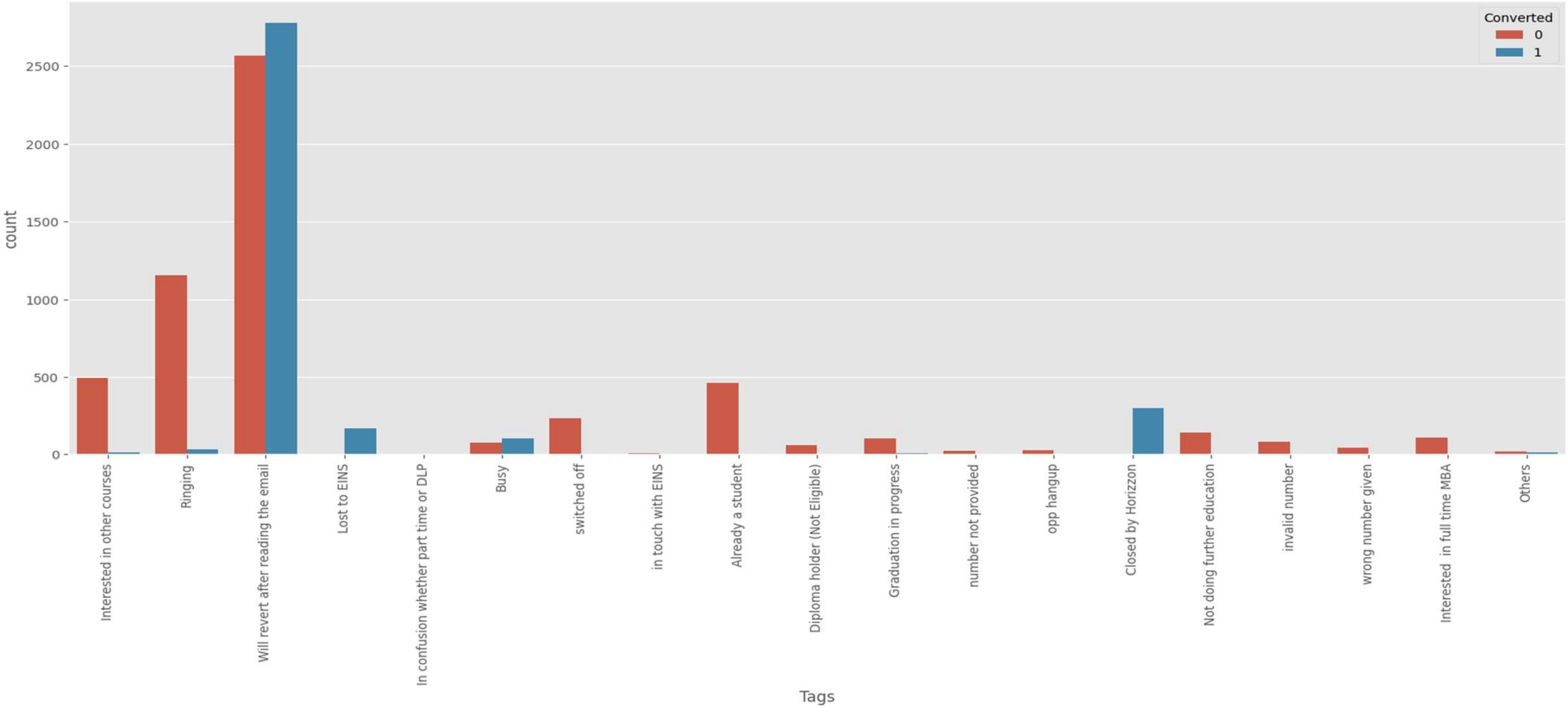
Current Occupation





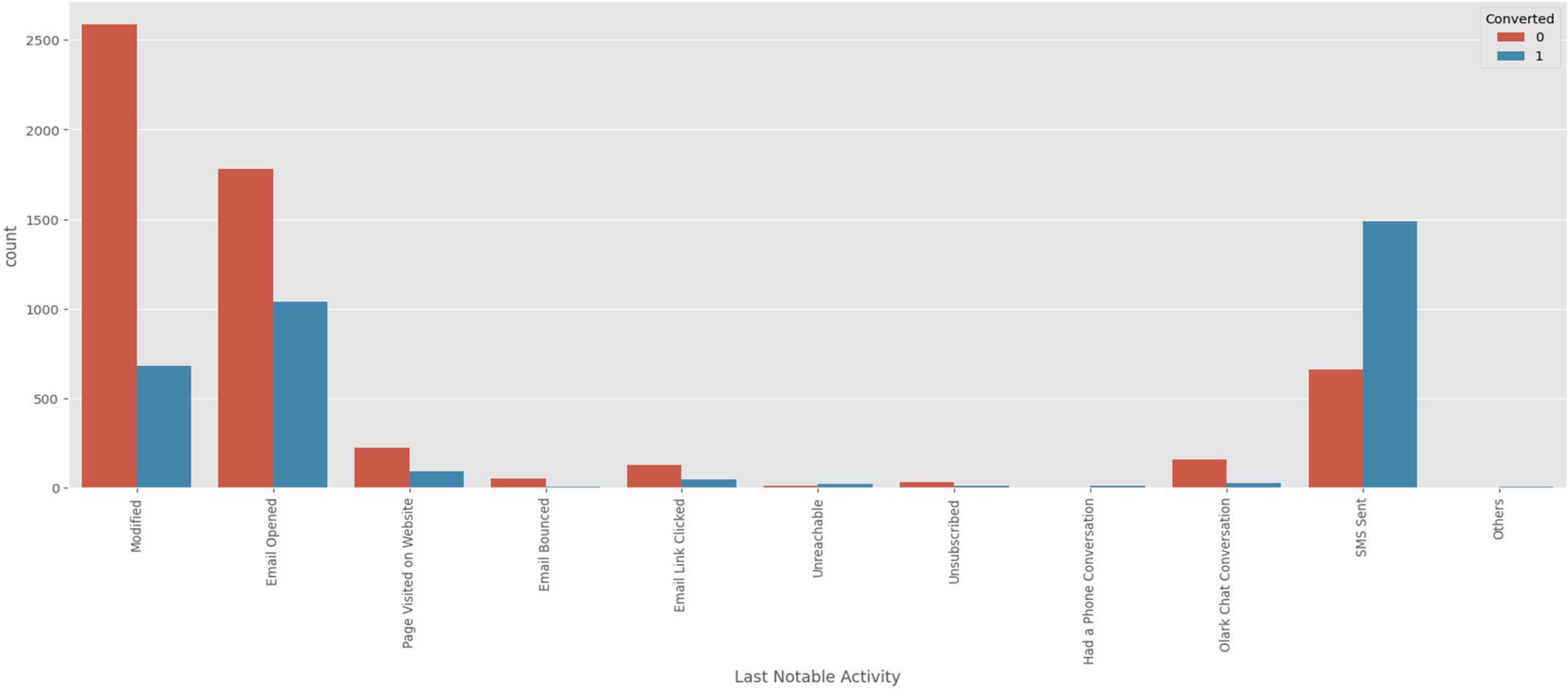
Working Professionals are most likely to get converted.

Tags



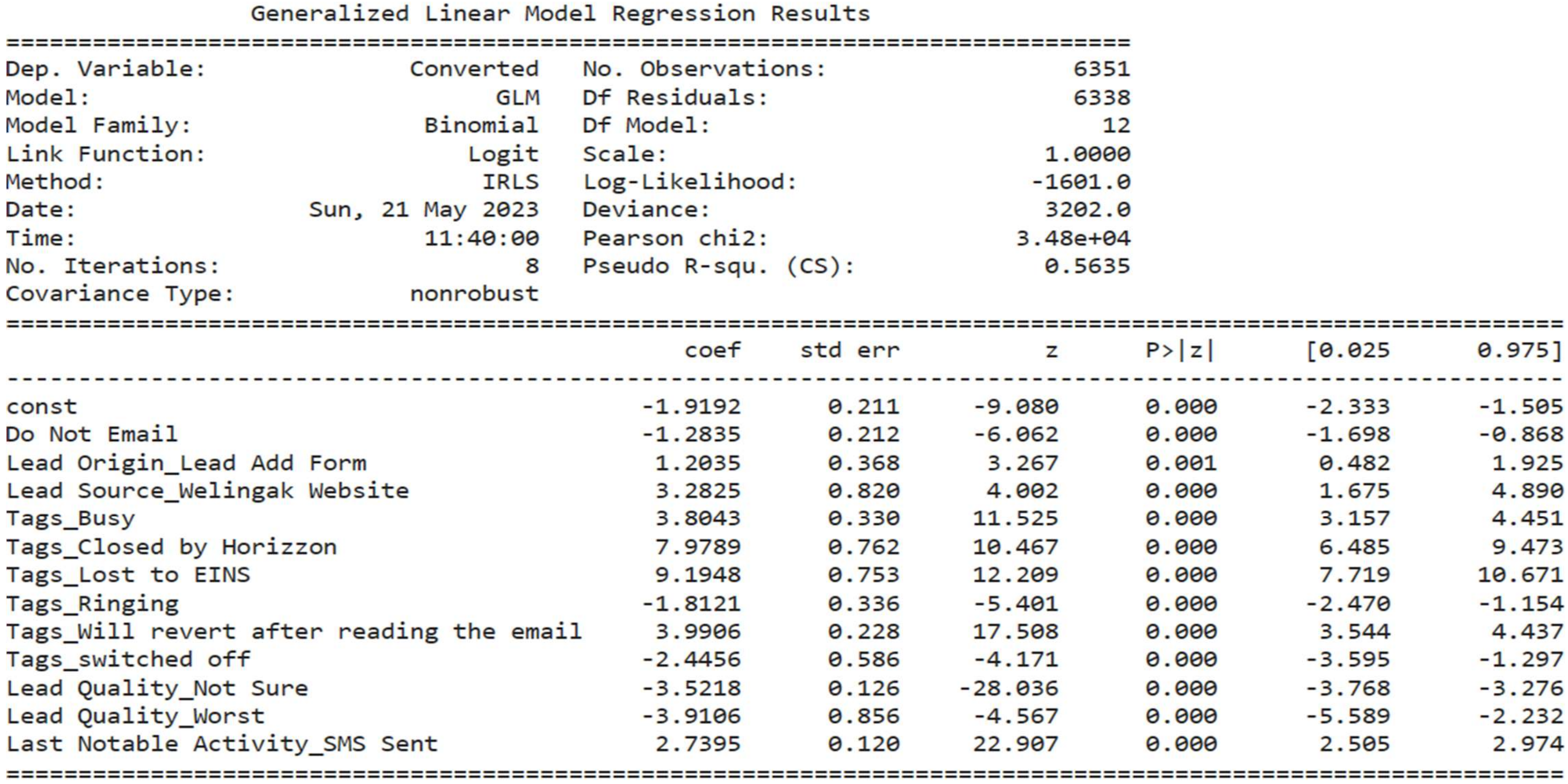
High conversion rates for tags ‘**Will revert after reading the email**’, ‘**Closed by Horizon**’, ‘**Lost to EINS**’, and ‘**Busy**’.

Last Notable Activity



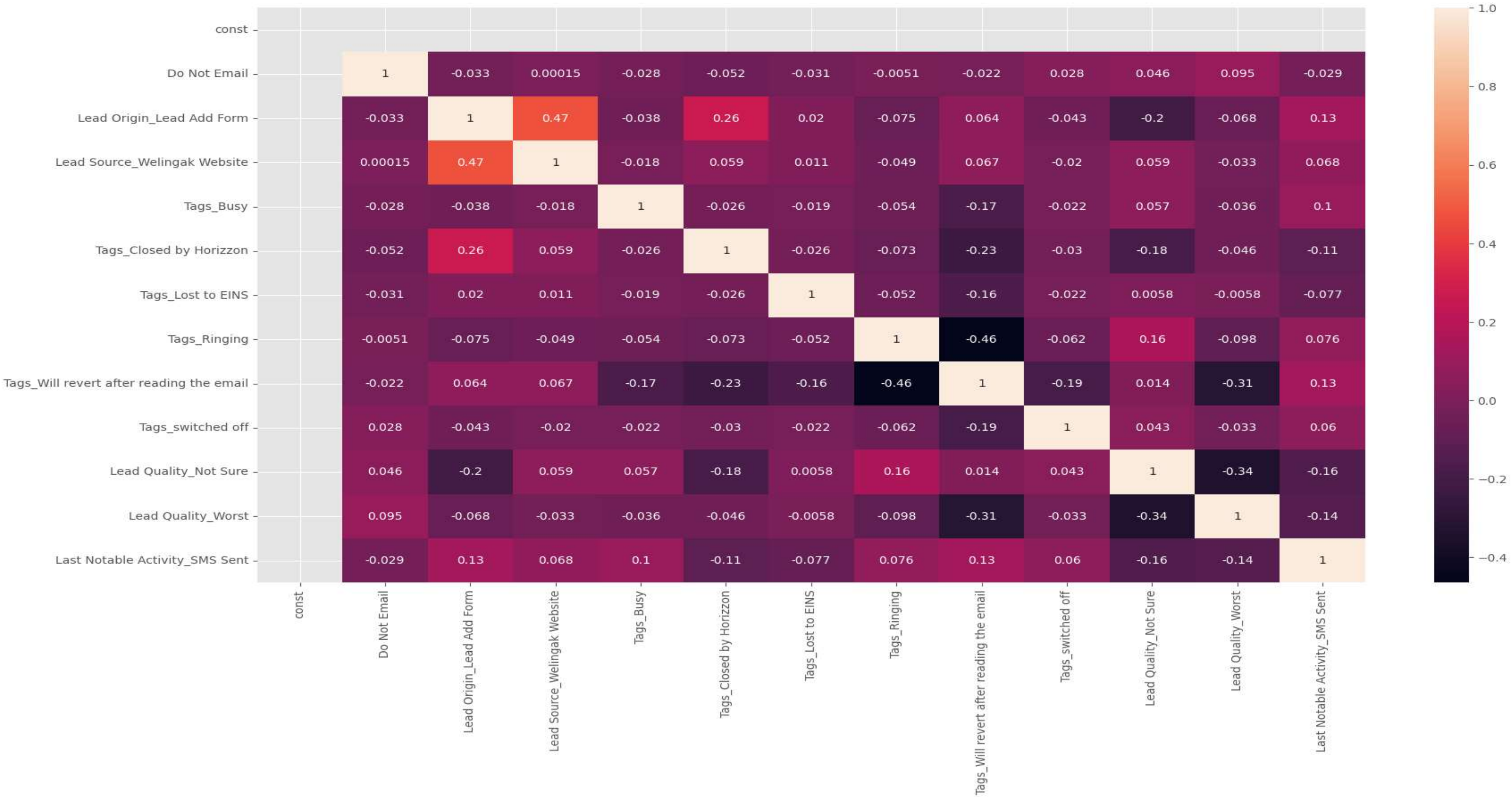
Highest conversion rate is for the last notable activity ‘**SMS Sent**’.

MODEL EVALUATION

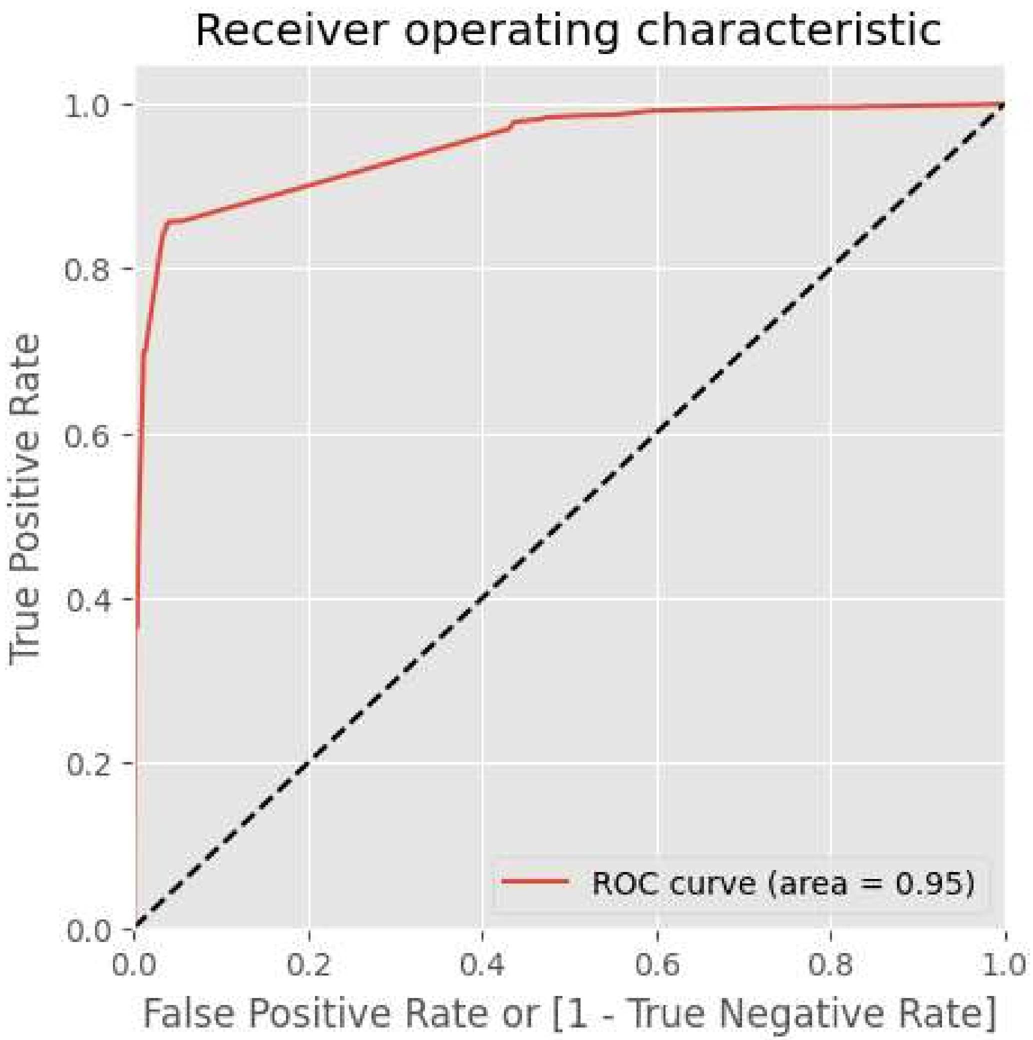


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6338
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1601.0
Date:                Sun, 21 May 2023   Deviance:                       3202.0
Time:                        11:40:00   Pearson chi2:                 3.48e+04
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5635
Covariance Type:            nonrobust
==============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                      -1.9192      0.211     -9.080      0.000      -2.333      -1.505
Do Not Email                               -1.2835      0.212     -6.062      0.000      -1.698      -0.868
Lead Origin_Lead Add Form                   1.2035      0.368      3.267      0.001       0.482       1.925
Lead Source_Welingak Website                3.2825      0.820      4.002      0.000       1.675       4.890
Tags_Busy                                   3.8043      0.330     11.525      0.000       3.157       4.451
Tags_Closed by Horizzon                     7.9789      0.762     10.467      0.000       6.485       9.473
Tags_Lost to EINS                           9.1948      0.753     12.209      0.000       7.719      10.671
Tags_Ringing                               -1.8121      0.336     -5.401      0.000      -2.470      -1.154
Tags_Will revert after reading the email    3.9906      0.228     17.508      0.000       3.544       4.437
Tags_switched off                          -2.4456      0.586     -4.171      0.000      -3.595      -1.297
Lead Quality_Not Sure                      -3.5218      0.126    -28.036      0.000      -3.768      -3.276
Lead Quality_Worst                         -3.9106      0.856     -4.567      0.000      -5.589      -2.232
Last Notable Activity_SMS Sent              2.7395      0.120     22.907      0.000       2.505       2.974
==============================================================================================================
```

Final Model Summary: All p-values are zero

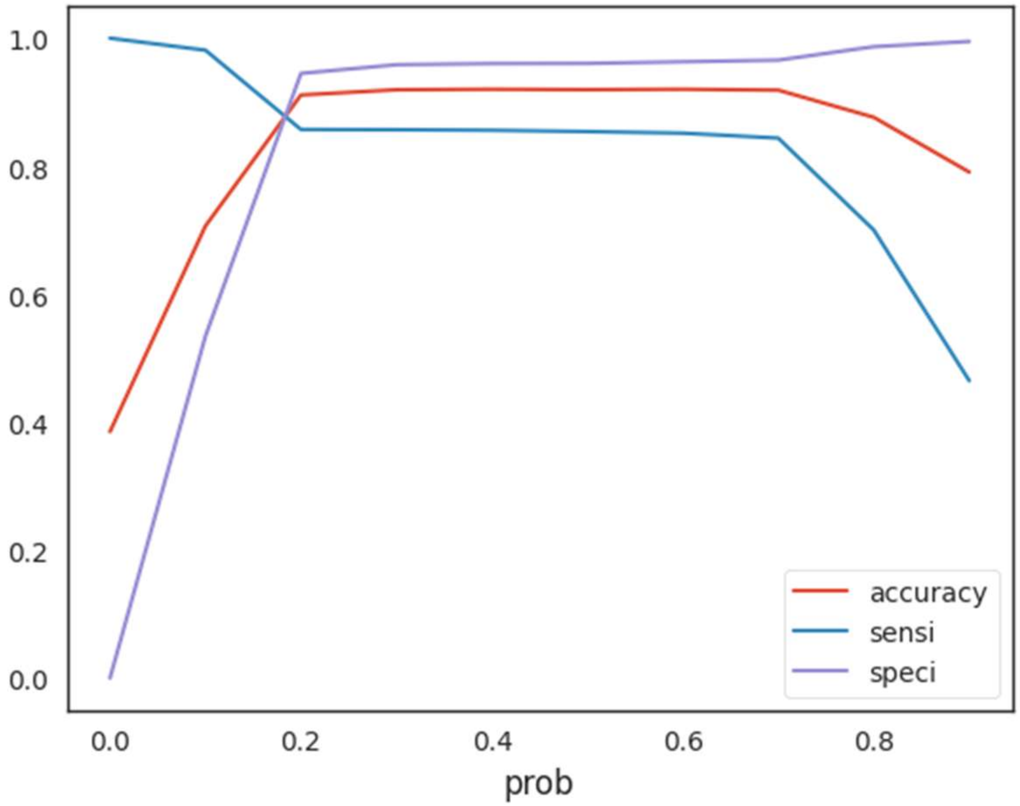


Correlations between features in the final model are **negligible.**



Area under curve = 0.95

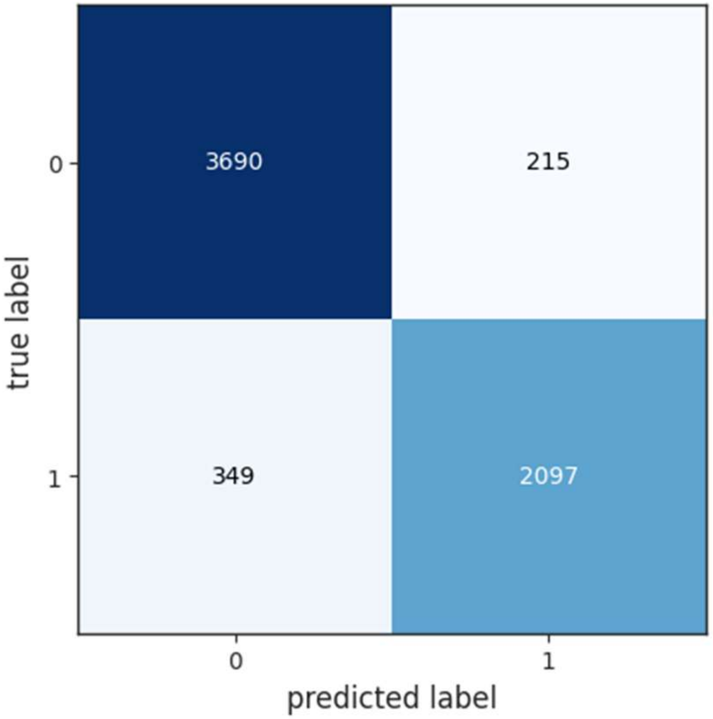
Finding Optimal Threshold



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

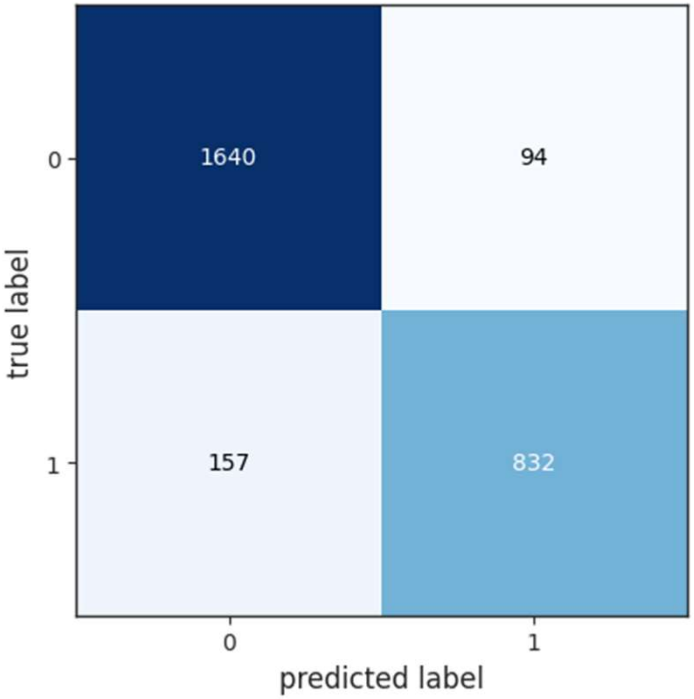
Optimal cutoff = 0.20

Confusion Matrix



true label \ predicted label	0	1
0	3690	215
1	349	2097

For train set



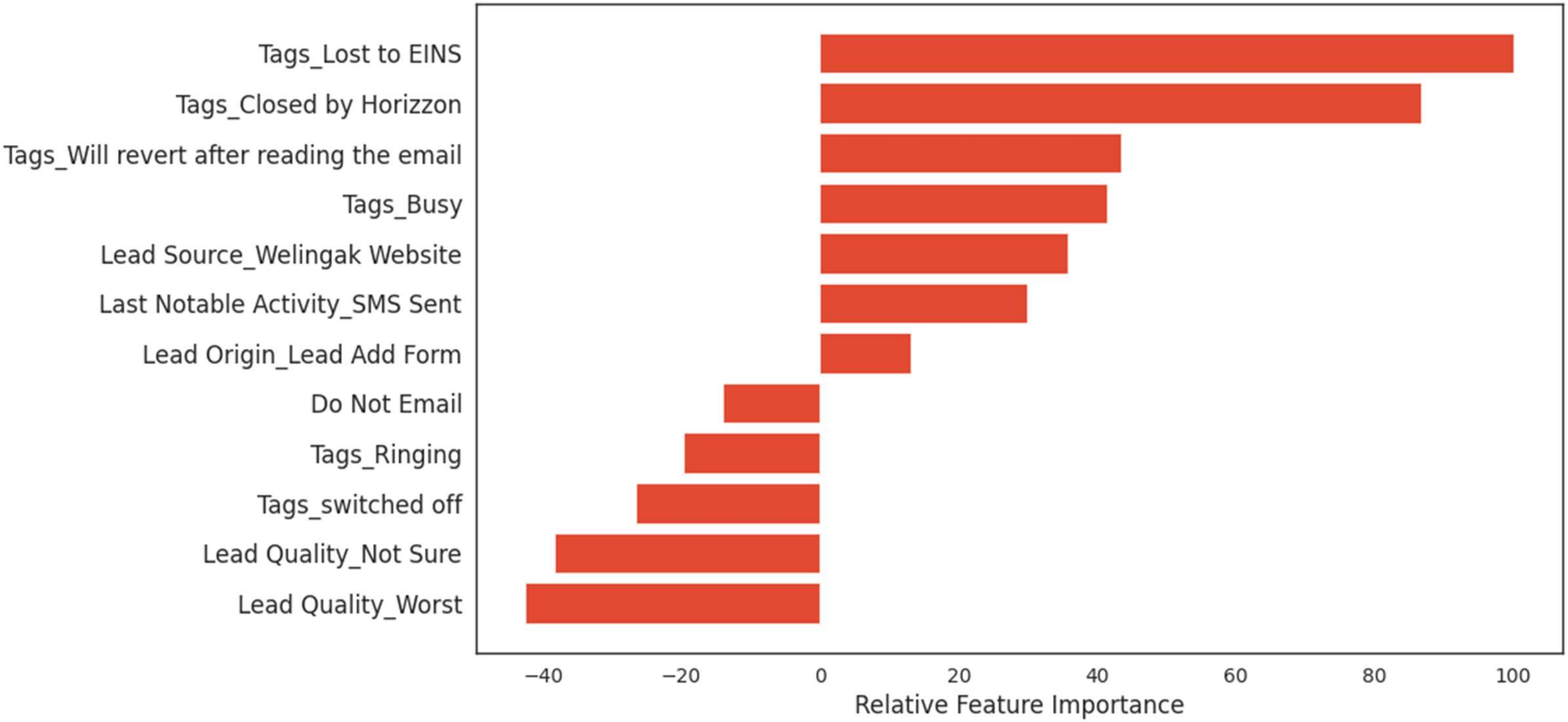
true label \ predicted label	0	1
0	1640	94
1	157	832

For test set

Final Results

Data	Train set	Test set
Accuracy	**0.9111**	**0.9078**
Sensitivity	0.8573	0.8412
Specificity	0.9449	0.9457
False Positive Rate	0.0550	0.0542
Positive Predictive Value	0.9070	0.8984
Negative Predictive Value	0.9135	0.9126
AUC	0.9488	0.9388

Relative Importance Of Features



INFERENCES

Feature Importance

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impactare:
  - ***Tags_Lost to EINS***
  - **Tags_Closed by Horizzon**
  - **Tags_Will revert after reading the email**
- These are dummy features created from the categorical variable Tags.
- All three **contribute positively** towards the probability of a lead conversion.
- These results indicate that the company should **focus more on the leads with these three tags**.

Recommendations

❑ By referring to the data visualizations, focus on

- *Increasing the conversion rates for the categories generating more leads and*
- *Generating more leads for categories having high conversion rates.*

❑ Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.

❑ Based on varying business needs, modify the probability threshold value for identifying potential leads.

THANK YOU